Illinois Institute of Technology

CSP 554 – Big Data Technologies

Project Draft

To Build a ETL Data Pipeline using Big Data SQL Technologies on Google Cloud Platform.

Name – Atharva Milind Nirali (A20517247)

Area of Project - Formula 1 / Auto Racing Sport.

- **Problem Statement / Introduction –**

  To build a data processing pipeline using Big Data technology to ingest, store, clean and use various data techniques to find out interesting insights / analytics from the Formula 1 Dataset using Google Cloud Platform.

- **What is Google Cloud Platform?**

The Google Cloud Platform is a set of cloud computing services that Google provides and is based on the same internal infrastructure as Google's consumer products like Google Search and YouTube. It offers several modular cloud services, such as computing, data storage, analytics, and machine learning, in addition to a set of management tools.

- **Overview of the project –**

1) Use of Formula 1 Dataset 1950 – 2022 to explore interesting insights.
2) Use Google Cloud to create a Data Pipeline to carry out ETL Processing using Cloud Storage, Data Flow and Big Query.
3) Use Data/Looker Studio (By Google) to visualize the insights gained from the F1 Dataset after careful analysis using Big Query.

Project Plan:

- **Extract (from Cloud Storage)**

Cloud Storage - Data can be stored and accessed on the Google Cloud Platform infrastructure using Google Cloud Storage, a RESTful online file storage web service. The service combines cutting-edge security and sharing features with the performance and scalability of Google's cloud.

- We store the CSV files of the Dataset on the Cloud Storage.
- Next, we store the Schema file on the Cloud Storage, this file basically defines the schema of the Big Query tables.
- Next, we store the transformation Function using which contains the transformation logic, on the cloud storage.

- **Transform and Load (using DataFlow)**

DataFlow - Dataflow is a fully managed data processing service for executing a wide variety of data processing patterns. It provides unified stream and batch data processing that's serverless, fast, and cost-effective.

- The DataFlow will execute the transformation, manage the parallel processing, co-ordinate the workers etc.
- Basically, DataFlow allow us to focus on the logic by taking care of the execution by itself.
- It uses the Apache Beam pipeline.
- Dataflow is responsible for the execution, i.e., it will take the input from the Cloud Storage, apply transformation as specified in the transformation function and then output the transformed data onto the BigQuery table using the schema specified.

- **Analyze (BigQuery)** 

BigQuery - We may manage and analyze data with the aid of BigQuery, a fully managed enterprise data warehouse. We can query terabytes in a matter of seconds and petabytes in a matter of minutes thanks to BigQuery's scalable, distributed analytical engine.

- BigQuery will be the target system for our data pipeline built on Google cloud platform.
- Tables created in the BigQuery will be loaded by a DataFlow job.
- We can also use these tables further analyze the data using SQL queries and store the output in the form on views.

- **Visualize (Data/Looker Studio by Google)** 

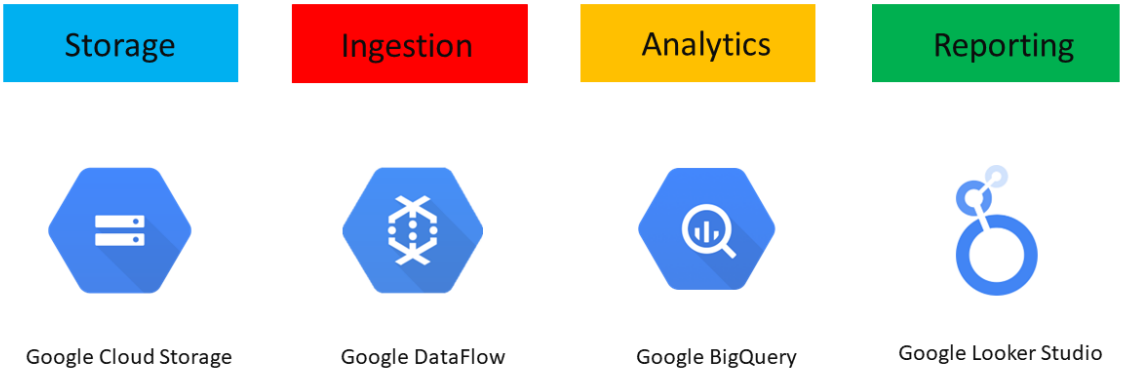
Looker - Looker is a powerful business intelligence (BI) tool that can be used to develop insightful visualizations. It offers a user-friendly workflow, is completely browser-based (eliminating the need for desktop software) and facilitates dashboard collaboration. Among other benefits, we can create interactive and dynamic dashboards, schedule and automate the distribution of reports, set custom parameters to receive alerts, and utilize embedded analytics.

- All we must do is specify the data source (Big Query Tables/Views) to the created reports on looker and it will automatically populate the results.

Results:

- We will have a complete Google Cloud Platform ETL Data Pipeline using the F1 Dataset.
- The Analysis will be reported/ visualized in the form or charts, plots, graphs etc. using Looker Studio.

ETL Data Pipeline using Google Cloud



References:

- https://www.kaggle.com/datasets/harrybassi13/formula-1
- https://cloud.google.com/bigquery/docs
- https://developers.google.com/looker-studio
- https://cloud.google.com/dataflow
- https://cloud.google.com/storage/docs
- https://cloud.google.com/dataflow/docs/quickstarts/create-pipeline-python#local-terminal
- https://cloud.google.com/dataflow/docs/guides/templates/provided-batch#cloud-storage-text-to-bigquery
- https://cloud.google.com/dataflow/docs/guides/data-pipelines
- https://cloud.google.com/storage/docs/creating-buckets
- https://www.cloudskillsboost.google/focuses/3460?parent=catalog